

Statistical Thermodynamics of Clustered Populations

Themis Matsoukas*

Department of Chemical Engineering, Pennsylvania State University, University Park, PA 16802

(Dated: August 19, 2014)

We present a thermodynamic theory for a generic population of M individuals distributed into N groups (clusters). We construct the ensemble of all distributions with fixed M and N , introduce a selection functional that embodies the physics that governs the population, and obtain the distribution that emerges in the scaling limit as the most probable among all distributions consistent with the given physics. We develop the thermodynamics of the ensemble and establish a rigorous mapping to regular thermodynamics. We treat the emergence of a so-called “giant component” as a formal phase transition and show that the criteria for its emergence are entirely analogous to the equilibrium conditions in molecular systems. We demonstrate the theory by an analytic model and confirm the predictions by Monte Carlo simulation.

PACS numbers: 02.50.Ey, 87.23.Cc 05.70.Ln 68.43.De

Keywords: Statistical thermodynamics, ensemble theory

I. INTRODUCTION

Gibbs’s ensemble method [1] represents a remarkable success of mathematical physics: it provides the complete macroscopic description of a complex stochastic system and prescribes its conditions of equilibrium and stability in terms of a small number of variables. The phenomenology of thermodynamics has proven useful in other areas. Systems of evolving discrete populations present a close analogy to thermodynamic microstates and have led to mathematical treatments that borrow from the language of equilibrium thermodynamics [2–6]. Polymer gelation [7–9], shattering in fragmentation [10], the spread of epidemics [11–13] and the emergence of connectivity in artificial and neural networks [14–17] are all systems that involve the emergence of a giant coherent structure, a process that is commonly discussed in the phenomenology of phase transitions. These diverse physical problems point to a common, yet unclear, link to thermodynamics. Jaynes drew a powerful connection between statistical mechanics and information theory [18] that suggests a broader applicability of thermodynamics. Jaynes considered the problem of a random variable whose probability distribution cannot be accessed except through limited data. He suggested that the unknown distribution can be identified as the one that is most probable among all distributions that are consistent with the available data. This amounts to a constrained maximization of the entropy of the distribution. Given appropriate constraints, practically any distribution may be interpreted as a maximum entropy distribution. The success of the maximum entropy formalism suggests that thermodynamic concepts could be extended outside the realm of interacting particles to any physical problem that involves a distributed population. The unknown distribution would then be determined by constrained maximization of en-

tropy. There is a problem however: there is no obvious connection between the constraints that are required to produce a maximum entropy distribution and the physics from which that distribution emerges. For example, depending on the rate law, particle growth by surface deposition may produce a distribution that is Gaussian, if the rate is independent of size, or exponential, if the rate is proportional to particle mass [19]. Both distributions can be derived by maximum entropy arguments, constrained either by fixing the mean (exponential), or the mean and the variance (Gaussian); there is no physical link, however, between these particular constraints and the rate laws from which they arise.

Here we develop a thermodynamic theory that circumvents these difficulties and connects a generic population to the laws that govern its evolution. Our approach is motivated by Jaynes in that we view the distribution in the scaling limit as the one that is most probable among all that satisfy the physics of the problem. We depart from Jaynes in that the distributions of the ensemble are not equally probable but biased by a *selection functional*, a quantity that embodies the physics under which the population evolves. The paper is organized as follows. In section II we define the microcanonical cluster ensemble and introduce the selection bias and its mathematical properties. In section III we pass to the thermodynamic limit, obtain the most probable distribution and derive the fundamental relationships among the primary variables of the ensemble. In section IV we discuss the analogies between the cluster ensemble and the familiar thermodynamic ensemble. In section V we develop the thermodynamic criterion for the emergence of a giant cluster (“gel”) and demonstrate the theory with an example, which we solve analytically and test by stochastic simulation. In section VI we discuss the findings in the broader context of stochastic processes and finally summarize the conclusions in section VII.

* matsoukas@engr.psu.edu

II. MICROCANONICAL CLUSTER ENSEMBLE

We begin with a population of M indistinguishable individuals divided into N distinguishable clusters ($N < M$). We define a configuration of masses to be an ordered list of N clusters with total mass M and notate it in vector form as $\mathbf{m} = (m_1, m_2, \dots, m_N)$. A configuration is characterized by its distribution $\mathbf{n} = (n_1, n_2, \dots)$, such that n_i is the number of clusters with i members. We refer to i as the size or mass of the cluster. The microcanonical ensemble consists of all possible configurations of M members are divided into N clusters. All distributions of the ensemble satisfy the constraints

$$\sum_i n_i = N, \quad \sum_i i n_i = M. \quad (1)$$

The maximum possible cluster size in a distribution of the (M, N) ensemble is $M - N + 1$, but if we adopt the convention $n_i = 0$ for all $i > M - N + 1$, the limits in the summations may be assumed to run from $i = 1$ to ∞ and will not be explicitly shown. Each distribution is associated with a multiplicity factor that represents the number of configurations that have the same distribution. This is equal to the number of permutations in the order of cluster masses in the configuration and is given by the multinomial coefficient,

$$\mathbf{n}! = \frac{N!}{n_1! n_2! \dots}. \quad (2)$$

The log of the multiplicity factor is the entropy of distribution:

$$S = \log \mathbf{n}!. \quad (3)$$

In the Stirling approximation this reverts to the familiar functional. We now introduce the *selection bias* $W(\mathbf{n})$, a functional of \mathbf{n} that biases selection such that the probability of distribution \mathbf{n} is proportional not only to its multiplicity $\mathbf{n}!$ but also to its selection bias $W(\mathbf{n})$:

$$P(\mathbf{n}) = \mathbf{n}! \frac{W(\mathbf{n})}{\Omega_{M,N}}. \quad (4)$$

Here $\Omega_{M,N}$ is the partition function and satisfies the normalization condition,

$$\Omega_{M,N} = \sum_{\mathbf{n}} \mathbf{n}! W(\mathbf{n}), \quad (5)$$

with the summation taken over all distributions of the ensemble. The selection bias is a fundamental property of the cluster ensemble and embodies the physics of the problem. Here, the bias will remain general and unspecified; the only condition we impose is that it must produce distributions with proper extensive behavior in the thermodynamic limit. This requires $\log W$ to be homogeneous in \mathbf{n} with degree 1 (see the Appendix for details).

It then follows from Euler's theorem that $\log W$ is a linear combination of its derivatives with respect to n_i :

$$\log W(\mathbf{n}) = \sum_i n_i \left(\frac{\partial \log W(\mathbf{n})}{\partial n_i} \right)_{n_j} \equiv \sum_i n_i \log w_i. \quad (6)$$

The derivatives $\log w_i$ are an important element of the theory; they represent the contribution of cluster size i to $\log W$ and will be referred to as *cluster bias*. If the log of the selection bias is a *linear* functional of \mathbf{n} [20], then the w_i are intrinsic functions of cluster size i , and thus independent of the distribution itself. A special case of linear bias is $W(\mathbf{n}) = 1$ ($\log w_i = 0$), which gives equal weight to all distributions (unbiased ensemble). In the general case the w_i 's will depend not only on i but also on the distribution \mathbf{n} to which they refer, as Eq. (6) implies.

The microcanonical ensemble defined here is closely related to that discussed in Pitman [21] and in Berestycki and Pitman [3] in the context of integer partitions. The main difference is in the natural multiplicity of distributions in the ensemble, which depends on whether cluster configurations are taken to be ordered (as in the cluster ensemble), or not (as in Refs. [3, 21]). Such differences can be reconciled between the two ensembles. References [3, 21] also present solutions for special cases of linear bias functionals that arise in coagulation and fragmentation. Here our interest is not in specific solutions but rather in the fundamental relationships between variables of the cluster ensemble.

Exchange Reactions

Suppose that two clusters in distribution \mathbf{n} with sizes i and j , respectively, exchange members to produce two new clusters with sizes k and l such that $i + j = k + l$. This binary exchange process converts one distribution of the ensemble into another and can be represented by the reversible reaction,



If the equilibrium constant is set to

$$K_{\mathbf{n}=\mathbf{n}'} = \frac{W(\mathbf{n}')}{W(\mathbf{n})}, \quad (8)$$

the process will then produce an ensemble of distributions whose probability is given by Eq. (4). The exchange reaction, allows us to associate the cluster ensemble, a static collection of distributions, with a dynamic ensemble that obeys detailed balance with respect to exchange reactions. The practical implication is that we may sample the ensemble by Metropolis Monte Carlo simulation of binary exchange reactions with acceptance probability proportional to $W(\mathbf{n}')/W(\mathbf{n})$.¹

¹ This extends to any type of exchange reaction, not necessarily binary with respect to clusters, as long as both M and N are

III. THERMODYNAMIC LIMIT

We now pass to the thermodynamic limit on the premise that when M and N are large the ensemble reduces to a single distribution, \tilde{n} . From Eq. (5) we then have

$$\log \Omega \rightarrow \tilde{S} + \log \tilde{W}, \quad (9)$$

where $\tilde{S} = S(\tilde{\mathbf{n}})$ is the entropy of the most probable distribution and $\log \tilde{W} = \log W(\tilde{\mathbf{n}})$ is its log-bias. The most probable distribution is obtained by standard Lagrange maximization, and the result is

$$\frac{\tilde{n}_i}{N} = \tilde{w}_i \frac{e^{-\beta i}}{q}, \quad (10)$$

where $\log \tilde{w}_i$ is the cluster bias in the most probable distribution. The Lagrange multipliers β and $\log q$ correspond to the two constraints in Eq. (1). They are given in terms of the ratio M/N in implicit form by

$$\frac{M}{N} = \sum i \tilde{w}_i e^{-\beta i} / \sum \tilde{w}_i e^{-\beta i}, \quad (11)$$

$$q = \sum \tilde{w}_i e^{-\beta i}, \quad (12)$$

which we obtain by inserting the most probable distribution into the constraints. We now take the log of the most probable distribution, multiply by \tilde{n}_i , and perform the summation (see Appendix):

$$\log \Omega_{M,N} = \beta M + (\log q) N, \quad (13)$$

This result of remarkable simplicity represents a fundamental relationship between the primary variables of the ensemble. Since S and $\log W$ are extensive in M and N , it follows from Eq. (9) that so is $\log \Omega$. In other words, $\log \Omega$ is homogeneous in M and N , with degree 1. By Euler's theorem we have

$$\beta = \left(\frac{\partial \log \Omega}{\partial M} \right)_N; \quad \log q = \left(\frac{\partial \log \Omega}{\partial N} \right)_M, \quad (14)$$

and

$$d \log \Omega_{MN} = \beta dM + (\log q) dN. \quad (15)$$

Equation (14) identifies β and q as the partial derivatives of the microcanonical log-partition function with respect to its extensive variables, while Eq. (15) governs the evolution of a population under a quasistatic change of state (dM, dN).

conserved. Binary reactions are convenient because they are the simplest to implement by simulation.

Canonical Ensemble

To complete the theory we derive the statistics of the canonical ensemble. We start with a large microcanonical ensemble of M' individuals distributed into N' clusters. The ensemble is characterized by fixed β, q . We sample randomly N clusters ($N \ll N'$) from a configuration of this microcanonical pool and seek the probability of distribution \mathbf{n} that is sampled in this manner. In the thermodynamic limit the probability to pick a cluster of size i is given by the relative frequency of cluster size i ensemble, $P_i = w_i \exp(-\beta i)/q$. For small N relative to N' the probability to sample distribution $\mathbf{n} = (n_1, n_2, \dots)$ is $\mathbf{n}! P_1^{n_1} P_2^{n_2} \dots$. Upon expanding the product the result becomes

$$P(\mathbf{n}) = \mathbf{n}! W(\mathbf{n}) \frac{e^{-\beta M}}{q^N}, \quad (16)$$

where M is the total mass (number of members) in the sampled distribution. Strictly, this derivation applies to linear bias because it assumes the w_i in Eq. (16) to be the same for all distributions. Nonetheless, a sharply peaked ensemble in the vicinity of the most probable distribution is adequately described by a linearized bias and the result applies to general bias in the scaling limit.

The canonical partition function is the sum of the canonical weights $\mathbf{n}! W(\mathbf{n}) \exp(-\beta M)$. Since probabilities in Eq. (16) are properly normalized, by summation over all \mathbf{n} on both sides we obtain

$$Q = q^N. \quad (17)$$

Here we recognize $\log q$ as the intensive form of the logarithm of the canonical partition function, $(\log Q)/N$.

IV. CONNECTION TO THERMODYNAMICS

An analogy to molecular thermodynamics now emerges. Compare Eq. (15) with the familiar thermodynamic relationship in the nVE ensemble,

$$\frac{dS}{k} = \frac{dE}{kT} + \frac{p dV}{kT} - \frac{\mu dn}{kT}.$$

In both cases, on the left-hand side we have the differential of a quantity whose maximization defines the equilibrium state, expressed on the right-hand side in terms of a set of extensive variables that fix the macroscopic state of the system. A one-to-one correspondence can be drawn between the two ensembles, as shown in Table I (to complete the analogy we must also set $\mu = 0$). Notice that the maximized quantity of the cluster ensemble is not the entropy of distribution, as in statistical mechanics, but the microcanonical partition function. In statistical mechanics the selection bias is set by the postulate of equal a priori probabilities. This corresponds to uniform bias $W = 1$ in our theory. Equation (9) then gives $\log \Omega = S$

TABLE I. Correspondence between properties of the cluster ensemble and thermodynamics. Other mappings are possible.

cluster ensemble	Thermodynamics
$\log \Omega$	$\rightarrow (\text{thermodynamic Entropy})/k$
M	\rightarrow Energy
β	$\rightarrow 1/kT$
N	\rightarrow Volume
$\log q$	$\rightarrow p/kT$

and thus we recover the familiar nVE ensemble. In general, $\log \Omega$ and entropy in the cluster ensemble are distinct properties, and between the two, it is the partition function that is of fundamental importance.

The thermodynamic mapping shown here is not unique as one may choose to associate β with $-\mu/kT$, for example; this mapping is implied in Ziff et al. [22], who refer to the quantity $e^{-\beta}$ as “fugacity” (symbol ξ in the original reference). In this sense, the correspondence between variables of the cluster ensemble and thermodynamics should not be taken literally but ought be viewed as a mathematical mapping that allows one to obtain relationships between variables of the cluster ensemble by translation of the corresponding result in thermodynamics. For example, by direct analogy to the thermodynamic result, $C_V > 0$, we may write, $(\partial M/\partial N)_\beta < 0$, a result that we can independently obtain by stability analysis on $\log \Omega$. Still, the mapping in Table I has a certain intuitive appeal. For example, a process may be viewed as “compression” if it causes the average size to increase (we imagine N to decrease at constant M), or “expansion,” if the average size decreases.

V. A PHASE TRANSITION: THE GIANT CLUSTER

If populations are thermodynamic entities, do they undergo the equivalent of phase equilibrium? In the context of the cluster ensemble, a “phase” is a distinct distribution, and “phase equilibrium” refers to the coexistence of two separate distributions within the same population, such that they both remain stable under exchange of members. We address this problem by considering the emergence of a giant cluster, namely, the transition from a population of finite clusters to one that contains one cluster that is of the order of M . The emergence of the giant component has long invited an analogy to condensation, but now we are in position to make a rigorous connection.

When M individuals are placed into N clusters, the maximum possible cluster size is $i_{\max} = M - N + 1$. The size range $i_{\max}/2 < i \leq i_{\max}$ is special: it can accommodate *at most one cluster*; the mass balance is not satisfied otherwise. A cluster in this region contains a

finite fraction of the total mass (its size is of the the order of M) but a vanishingly small fraction of the total number of clusters. We refer to clusters in this range as the “gel” phase and to all others as “sol” phase. Such system distributes members between the two phases so that the partition function of the combined system is maximized under the constraints $M_{\text{sol}} + M_{\text{gel}} = M$ and $N_{\text{sol}} + N_{\text{gel}} = N$ with $N_{\text{gel}} = 1$. This variational principle leads to the condition

$$\left(\frac{\partial \log \Omega_{\text{sol}}}{\partial M_{\text{sol}}} \right)_{N_{\text{sol}}} = \left(\frac{\partial \log \Omega_{\text{gel}}}{\partial M_{\text{gel}}} \right)_{N_{\text{gel}}} \equiv \beta. \quad (18)$$

This result is obtained by maximizing the partition function of the two-phase system (see Appendix) and establishes the condition of thermal equilibrium between the phases. Since the number of clusters in each phase is fixed, q is *not* required to equilibrate. The emergence of a giant cluster is akin to osmotic equilibrium.

A Gelling Selection Bias

We demonstrate the application of the theory using a selection bias that can be solved analytically. We take the cluster bias to be $w_i = i^{-3}$, which corresponds to the selection functional,

$$W(\mathbf{n}) = \prod_i i^{-3n_i}.$$

Let us examine whether it is possible to construct a two-phase solution for this bias. The distribution of the sol phase is

$$\left(\frac{n_i}{N} \right)^{\text{sol}} = i^{-3} \frac{e^{-\beta i}}{q}, \quad (19)$$

with β and q given by

$$\frac{M}{N} = \frac{\text{Li}_2(e^{-\beta})}{\text{Li}_3(e^{-\beta})} \quad (20)$$

$$q = \text{Li}_3(e^{-\beta}). \quad (21)$$

where $\text{Li}_n(x)$ is the polylogarithm function. These are obtained From Eqs. (11) and (12) by letting the upper limit to go to infinity (scaling limit). To obtain β and q we solve Eqs. (20) and (21) numerically for the given M/N . The sol distribution is therefore completely defined if M/N is fixed.

Suppose the system forms a two-phase state that consists of a gel cluster with mass M_{gel} in a equilibrium with sol ($M_{\text{sol}} = M - M_{\text{gel}}$, $N_{\text{sol}} = N - 1 \rightarrow N$). For the linear bias functional, the partition function of the gel-phase is $\Omega_{\text{gel}} = w(M_{\text{gel}})$, where $w(M_{\text{gel}})$ is the cluster bias of the giant cluster. This follows from Eq. (9) and the fact that the gel phase consists of a single cluster ($S_{\text{gel}} = 0$). Its temperature is

$$\beta^{\text{gel}} = \left. \frac{d \log w_i}{d i} \right|_{M_{\text{gel}}} = -3/M_{\text{gel}}, \quad (22)$$

TABLE II. Phase behavior of linear cluster bias $w_i = i^{-3}$. The system forms a giant cluster at $M/N = 1.368$.

	$M/N < 1.368$	$M/N \geq 1.368$
	single sol	sol+gel
sol distribution	$\frac{n_i}{N} = \frac{e^{-\beta i}}{q} i^{-3}$	$\frac{n_i}{N} = 0.832 i^{-3}$
gel distribution	—	$n_i^{\text{gel}} = \delta_{i, M_{\text{gel}}}$
β	Eq. 20	0
q	Eq. 21	1.202
gel fraction	0	$\frac{M_{\text{gel}}}{M} = 1 - 1.368 \frac{M}{N}$

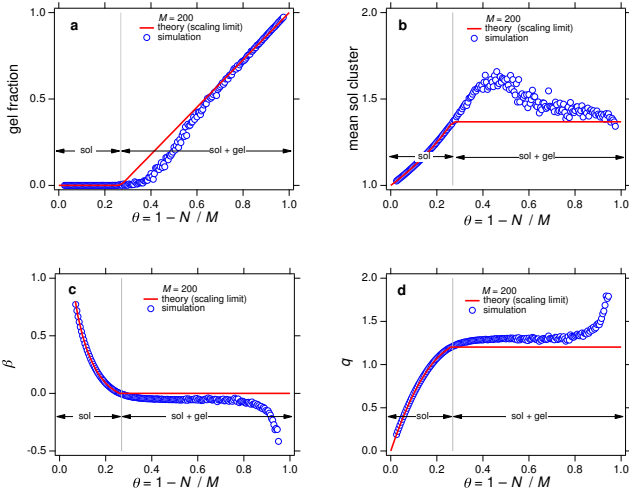


FIG. 1. (Color online) Phase diagram of a gelling system with cluster bias $w_i = i^{-3}$: (a) gel fraction, (b) mean sol cluster, (c) temperature β , and (d) pressure q , as a function of the progress variable $\theta = 1 - N/M$. The simulations are conducted with $M = 200$. The theoretical lines are based on the equations in Table II, which assume an infinite system.

and in the scaling limit, $\beta^{\text{gel}} \rightarrow 0$. The corresponding equilibrium sol phase is obtained from Eq. (19) with $\beta = 0$, $q = 1.202$. This produces the power-law distribution

$$\left(\frac{\tilde{n}_i}{N}\right)^* = 0.832 i^{-3}, \quad (23)$$

whose mean cluster size is $i^* = 1.368$. The phase behavior is now fully determined. For $M/N < i^* = 1.368$ we have a single sol phase whose distribution is,

$$\frac{n_i}{N} = \frac{i^{-3} e^{-i\beta}}{\text{Li}_3(e^{-\beta})}. \quad (24)$$

If M/N is larger than the equilibrium sol cluster, the system splits in two phases: a sol phase whose distribution

is given in Eq. (23), and one giant cluster whose fraction is given by the mass balance condition (see Appendix)

$$\phi_{\text{gel}} = 1 - i^* \left(\frac{N}{M}\right), \quad (25)$$

where $i^* = 1.368$ is the equilibrium mean sol cluster. Notice that the normalized equilibrium sol distribution $(n_i/N)_{\text{sol}}$ remains constant in the entire sol/gel region. The conversion of the sol phase into the gel along the path $N \rightarrow 1$ at fixed M occurs isothermally until the sol disappears completely.

We test these predictions by Monte Carlo simulation using the method of exchange reactions on a finite population with $M = 200$ members. We begin with a configuration (ordered list) of N cluster masses with total mass M ; the initial distribution is approximately uniform but any other distribution may be used. At each step we pick two clusters at random, merge them into a single cluster, then break it up randomly in two. The new configuration is accepted with probability proportional to the equilibrium constant in Eq. (8) and the process is repeated until the system reaches equilibrium. The mean distribution is calculated as an average of 1 to 4×10^6 realizations. The simulation is repeated with N ranging from $M - 1$ to 2.

First we analyze the sol and gel phases obtained in the simulation. We calculate the gel fraction, ϕ_{gel} , as the fraction of the total mass that resides in the region $i > (M - N + 1)/2$. The mean cluster size in the sol (\bar{i}) is calculated as the average mass in the region $1 \leq i < (M - N + 1)/2$, and the parameters β and q are obtained from the finite version of Eqs. (20) and (21). These results (ϕ , \bar{i} , β and q) are shown in Fig. 1 as a function of the progress variable $\theta = 1 - N/M$, where $\theta = 0$ refers to a fully dispersed population (perfect sol), and $\theta = 1$ a fully gelled population (all particles have joined the giant gel). The simulations are in general agreement with the theory though discrepancies are noted due to finite-size effects. The gel point is predicted at $\theta^* = 1 - 1/i^* = 0.2690$ ($N^* = 146$), while the simulation gives $\theta^* \approx 0.3$ ($N^* \approx 140$). The gel fraction is seen to be a linear function of θ with small deviations seen near the gel point. The mean cluster in the post-gel region converges to the predicted value $i^* = 1.368$; its small value makes it quite more sensitive to finite-size effects. Recall that in these simulations N is varied at constant M . In thermodynamics terms, in the direction of increasing θ the system undergoes compression. Accordingly its temperature $1/\beta$ increases, and so does its pressure $\log q$.

Sample distributions are shown in Fig. 2. At $N = 170 > N^*$ the population consists of a single sol distribution that decays monotonically within the sol region and is in excellent agreement with Eq. (19). At $N = 75 < N^*$ the population is compressed above its gel point, a gel phase is present, and the sol distribution is now given by Eq. (23). As N is reduced further, the sol phase “shrinks in place,” i.e., the number of sol

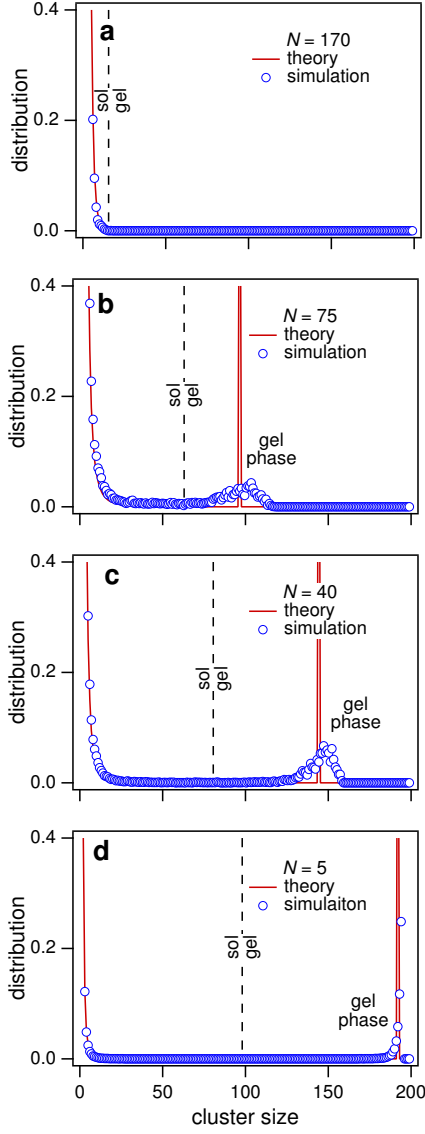


FIG. 2. (Color online) Selected distributions for cluster bias $w_i = i^{-3}$ in a finite population with $M = 200$ members. The giant cluster appears near $N^* = 146$. The shaded area is the simulated distribution and the solid line is theory. The dashed line marks the phase boundary between the sol phase and the gel phase.

clusters decreases while the normalized distribution remains fixed, and the giant cluster moves to larger size as more particles are converted from the sol. The theoretical distribution of the gel phase is a Kronecker delta at $M_{\text{gel}} = M(1 - 1.368 N/M)$. The apparent distribution of gel clusters represents not the relative number of gel clusters in a typical distribution (there is at most one cluster in this range) but fluctuations in the position of the giant cluster from one configuration to the next.

We examine these fluctuations in more detail in Fig. 3, which shows the mass fraction of the giant cluster,

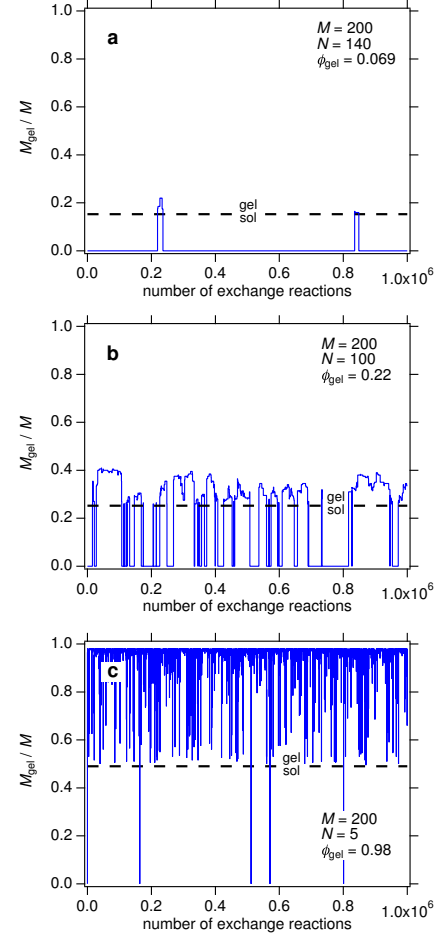


FIG. 3. (Color online) Fluctuations of the gel cluster with cluster bias $w_i = i^{-3}$ over the course of exchange reactions in a population with $M = 200$ members. (a) $N = 140$ (right about the gel point); (b) $N = 100$ (above the gel point); (c) $N = 5$ (very close to complete gelation). The dashed line marks the boundary between the sol and the gel phase. When the gel cluster falls below this line, it enters the sol phase and the gel fraction drops to zero.

M_{gel}/M , within individual configurations of the ensemble and how it evolves as a result of exchange reactions. The dashed line marks the sol-gel boundary. A cluster above this line is a gel cluster; if the maximum cluster falls below the dashed line (sol region), no gel cluster exists, i.e. the giant cluster evaporates into the sol and its mass is zero. Crossings above and below the dashed line indicate condensation/evaporation, respectively, of the gel cluster. These fluctuations therefore reflect transfer between the two phases. At $N = 140$ ($\theta = 0.3$) the system is close to the gel point. The nucleation of the gel phase is indicated by short-lived excursions into the gel region. The gel fraction, calculated over all states (whether they contain a gel cluster or not) is $\phi_{\text{gel}} = 0.069$. Even though the giant cluster appears abruptly in the size range $i > i_{\text{max}} = 30$, the ensemble-average gel frac-

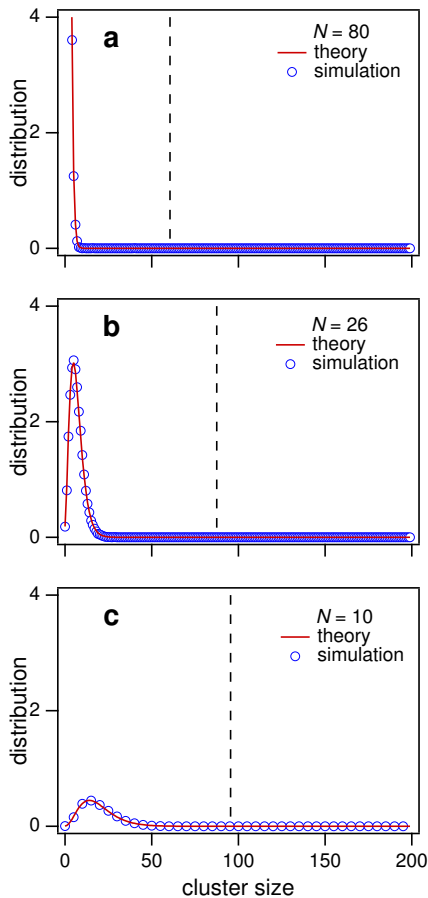


FIG. 4. (Color online) The cluster bias $w_i = i^3$ results in a stable single-phase population for all N (no gel phase). The shaded area is the simulated distribution and the solid line is theory. The dashed line marks the phase boundary between sol and the region of the giant cluster.

tion at the gel point is continuous. At $N = 100$ ($\theta = 0.5$) the system is above the gel point. Most states contain a gel cluster that makes frequent crossings into the sol region. The mean gel fraction is $\phi_{\text{gel}} = 0.22$. At $N = 4$ the system is very close to complete gelation with most of its mass residing on the giant cluster ($\phi_{\text{gel}} = 0.98$). Even at this highly gelled state, fluctuations reach the sol-gel boundary and occasionally the giant cluster evaporates but these excursions are both less frequent and short-lived.

The phase behavior of the gelling bias studied here bears remarkable similarity to Stockmayer's theory of gelation [23]. Stockmayer constructs an ensemble of polymer chains, similar to our cluster ensemble, and selects chains in proportion to the number of ways that i monomers can be combined to form an i -mer. He obtains the most probable distribution by maximizing the partition function and finds the gel fraction to be a linear function of the fraction of unreacted bonds, a quantity

analogous to our θ . While Stockmayer raises numerous analogies between his gelation model and the thermodynamics of phase transitions, he obtains phase diagram by analysis of the convergence radius of the cluster distribution, not by thermodynamics, and identifies the gel point as the limiting condition that guarantees convergence of the moments of order 0 and 1. We can now make the thermodynamic connection rigorous. The cluster bias in Stockmayer's model is² [23]

$$w_i = \frac{f!(fi - i)!}{i! (fi - 2i + 2)!}, \quad (26)$$

where f is the functionality of the monomer. For $f = 3$ (the general case $f > 3$ can be treated similarly) this is functionally equivalent to the cluster bias (see the Appendix)

$$w_i = i^{-5/2}. \quad (27)$$

This is of the same power-law type as the gelling bias discussed above and leads to the same isothermal condition at the sol-gel point, $\beta^{\text{gel}} = \beta^{\text{sol}} = 0$. This condition also defines the radius of convergence of the summations in Eqs. (11) and (12): the first and second moments of \tilde{n}_i converge in the region $\beta \geq 0$, while the second moment converges in $\beta > 0$ but diverges at the gel point. This is precisely the method by which Stockmayer obtained the gel point. Here we arrived at the same result by strict thermodynamic analysis.

Not every selection bias leads to a gelling population. Consider the cluster bias $w_i = i^3$, the inverse of the case discussed above, whose sol distribution is

$$n_i/N = i^3 e^{-\beta i}/q.$$

The temperature of the giant cluster, if one forms, is $\beta^{\text{gel}} = 3/M_{\text{gel}}$. Again, in the scaling limit we obtain $\beta^{\text{gel}} \rightarrow 0$ but now the corresponding sol distribution diverges and cannot produce a sol that satisfies the two constraints of the ensemble. It is not possible to construct an equilibrium two-phase mixture in this case. Even though this selection bias favors distributions with heavy tails, these are fully contained in the sol region. This prediction is confirmed by Monte Carlo simulation (Fig. 4), which demonstrates that the system forms a single sol distribution for all M/N .

VI. DISCUSSION

The main element of the theory is the selection bias W , which alone determines the distribution in the scaling limit. The connection to statistical mechanics may now be seen more clearly. In statistical mechanics we

² Notice that the w_i in Stockmayer's notation refers to $i!w_i$ in ours.

begin with the postulate of equal a priori probabilities, which corresponds to the unbiased ensemble with $W = 1$, and obtain the exponential distribution of microstates as the most probable distribution in the scaling limit. By allowing the selection bias to be an arbitrary functional of the distribution \mathbf{n} , it is possible to obtain *any* distribution in the scaling limit. Conversely, any distribution of M members clustered in N groups can be associated with an equilibrium ensemble that obeys thermodynamics, provided that the corresponding selection bias is identified. The broader implication is that stochastic variables may be treated in the formalism of thermodynamics by associating its probability density function with the corresponding equilibrium ensemble. In this sense, thermodynamics should be viewed as a probabilistic calculus that is applicable not only to systems of physical particles, but to stochastic processes in general. Complex problems such as gelation, percolation and network connectivity may then be treated as formal phase transitions that obey thermodynamics. The crux of the problem then is the determination of the selection bias. This must be done on a case-by-case basis, based on the physical laws that govern the system at hand. In the Stockmayer model, for example, the selection bias arises from a combinatorial model for polymer chains. In populations undergoing dynamic transformations, it is determined by the corresponding rate laws. Examples will be discussed elsewhere.

We close with a final note on dynamics. Although time is not an explicit element of the theory, the connection to dynamics is provided by Eq. (15). This equation describes the evolution of the ensemble along a quasi-static trajectory, which we define as a path along which the selection bias remains invariant, while M and N change due to dynamic processes in the population. Essentially, Eq. (15) gives the evolution of the system in terms of the mean cluster size M/N rather than time. At fixed (M, N) the system is at “equilibrium,” which is understood to mean that the ensemble relaxes to the most probable distribution that corresponds to the present values of M and N . Even if the system is on an irreversible path, i.e., unidirectional in time, the state at fixed (M, N) may be described using the calculus of equilibrium thermodynamics. The validity of this approach is confirmed in Fig. 1: in moving from $\theta = 0$ to $\theta = 1$ we follow a polymerization reaction that leads to the irreversible formation of a gel.

VII. CONCLUSIONS

In summary, we have presented a thermodynamic theory for a generic population of M individuals clustered into N groups. The distribution of this population is viewed as the most probable distribution that emerges among all possible partitions of M into N elements, under a functional that biases the selection of individual partitions. By associating the population with a corre-

sponding ensemble we are able to express the distribution in terms of thermodynamic quantities (partition function, “temperature,” “pressure”) that obey relationships analogous to those in molecular systems. An important result of the theory is a thermodynamic criterion (maximization of the partition function) to determine whether a giant cluster (gel phase) forms. Although it is common in the literature to refer to the emergence of a giant component as a phase transition, the theory presented here associates this process, for the first time, with the maximization of a thermodynamic functional. This unlocks the toolbox of thermodynamics and makes it available to the study of population balances and stochastic phenomena in general.

APPENDIX

Appendix A: Derivations and Additional Notes

1. Homogeneity Condition Eq. (6)

Here we show that in order for the ensemble to have proper extensive behavior in the thermodynamic limit, the log of the selection bias must be a homogeneous function of n_i with degree 1. Proper extensive behavior means that if M and N are both increased by a factor λ , the number of clusters \tilde{n}_i in the most probable distribution must increase by the same factor. Equivalently, the ratio \tilde{n}_i/N must be an intensive property, i.e., a function of M/N . The most probable distribution

We start with the relationship between β , q and M/N in Eqs. (11), (12), which we rewrite as

$$\frac{M}{N} = \sum_i i e^{-\beta i + \log \tilde{w}_i} / \sum_i e^{-\beta i + \log \tilde{w}_i}, \quad (\text{A1})$$

$$q = \sum_i e^{-\beta i + \log \tilde{w}_i}. \quad (\text{A2})$$

The ratio \tilde{n}_i/N will be intensive only if β , q , and $\log \tilde{w}_i$ are intensive; in turn, β and q will be intensive only if the $\log \tilde{w}_i$ are intensive. This implies that if a distribution \mathbf{n} is multiplied by λ , which will result in all cluster numbers being multiplied by λ while M/N remains constant, the cluster bias $\log w_i$ must remain unchanged:

$$\log w_i = \left(\frac{\partial \log W(\lambda \mathbf{n})}{\partial (\lambda n_i)} \right) = \left(\frac{\partial \log W(\mathbf{n})}{\partial n_i} \right).$$

This requires

$$\log W(\lambda \mathbf{n}) = \lambda \log W(\mathbf{n})$$

and states that $\log W(\mathbf{n})$ is homogeneous function of \mathbf{n} of degree 1.

2. Derivation of Fundamental Eq. (13)

Take the log of Eq. (10)

$$\log \frac{\tilde{n}_i}{N} = \log \tilde{w}_i - \beta i - \log q,$$

multiply both sides by \tilde{n}_i and sum over all i :

$$\sum \tilde{n}_i \log \frac{\tilde{n}_i}{N} = \sum \tilde{n}_i \log \tilde{w}_i - \beta M - (\log q)N,$$

Using $S = \log \tilde{\mathbf{n}}!$, and the Euler relationship in Eq. (6), we obtain

$$S + \log \tilde{W} = \beta M + (\log q)N,$$

which, combined with Eq. (9), leads to the final result

$$\log \Omega = \beta M + (\log q)N. \quad (\text{A3})$$

If we write this result as

$$\log \Omega = M \left(\beta + \frac{N}{M} \log q \right) = N \left(\frac{M}{N} \beta + \log q \right), \quad (\text{A4})$$

it will become clear that $\log \Omega$ is homogeneous of degree 1 with respect to both M and N (β , $\log q$ and M/N are all intensive), and satisfies Euler's theorem:

$$\log \Omega = M \left(\frac{\partial \log \Omega}{\partial M} \right) + N \left(\frac{\partial \log \Omega}{\partial N} \right). \quad (\text{A5})$$

Comparison with Eq. (A3) leads to

$$\beta = \left(\frac{\partial \log \Omega}{\partial M} \right)_N, \quad \log q = \left(\frac{\partial \log \Omega}{\partial N} \right)_M,$$

which appears as Eq. (14) in the text.

3. Derivation of Isothermal Condition Eq. (18)

At equilibrium, a sol-phase system maximizes the microcanonical weight of the composite system. For a given partitioning of mass between the two phases, the sol phase maximizes its own microcanonical weight when it relaxes to the most probable distribution that corresponds to $M_{\text{sol}} = M - M_{\text{gel}}$, $N_{\text{sol}} = N - 1$. Therefore, we seek to maximize the partition function of the combined system,

$$\log \Omega_{M,N} = \log \Omega_{M_{\text{sol}},N-1} + \log \Omega_{\text{gel}}.$$

Suppose that the giant cluster exchanges mass δm with the sol. Such exchange must leave the overall partition function unchanged:

$$0 = \delta \log \Omega_{M_{\text{sol}},N-1} + \delta \log \Omega_{\text{gel}}.$$

We divide both sides by δm , and noting that $\delta M_{\text{gel}} = -\delta M_{\text{sol}}$, the result is

$$\frac{\delta \log \Omega_{\text{sol}}}{\delta M_{\text{sol}}} = \frac{\delta \log \Omega_{\text{gel}}}{\delta M_{\text{gel}}}.$$

Since the exchange is at constant number of clusters, we identify these derivatives as the temperatures in each phase and thus arrive at the isothermal condition:

$$\beta_{\text{sol}} = \beta_{\text{gel}}.$$

Since the number of clusters does not fluctuate during exchange reactions, pressures do not equalize and each phase may be at its own pressure.

4. The Sol-Gel Boundary

Here we determine the cluster size that separates the sol from the gel. Suppose that i_{sol} is the *maximum possible* cluster size in the sol region and i_{gel} is the *minimum possible* cluster size in the gel region. Since i_{gel} and i_{sol} are neighboring masses, they meet the condition

$$i_{\text{gel}} - i_{\text{sol}} = 1. \quad (\text{A6})$$

If a mass i_{gel} is placed in the gel, we are left in the sol with $M_{\text{sol}} = M - i_{\text{gel}}$, $N_{\text{sol}} = N - 1$. Then the maximum possible cluster size in the sol is

$$i_{\text{sol}} = M_{\text{sol}} - N_{\text{sol}} + 1 = M - N - i_{\text{gel}} + 2. \quad (\text{A7})$$

Combining with Eq. (A6) we find

$$i_{\text{sol}} = \frac{M - N + 1}{2}, \quad i_{\text{gel}} = \frac{M - N + 3}{2}. \quad (\text{A8})$$

Therefore, we identify the sol and the gel regions by the conditions:

$$\begin{aligned} \text{sol} : & \quad 1 \leq i \leq i_{\text{max}}/2; \\ \text{gel} : & \quad i_{\text{max}}/2 < i \leq i_{\text{max}}, \end{aligned}$$

where $i_{\text{max}} = M - N + 1$ is the maximum possible cluster in the (M, N) ensemble.

5. Stockmayer's model

Here we derive the asymptotic form of Eq. (26) for trifunctional monomers. Using the Stirling approximation, $x! \approx x^x \exp(-x) \sqrt{2\pi x}$, and $f = 3$ the Stockmayer weight becomes

$$w_i = \frac{e^2}{\sqrt{\pi}} (12i)^i (i+2)^{-i-5/2}. \quad (\text{A9})$$

Noting that

$$(i+2)^{-i-5/2} \sim i^{-i-5/2} e^{-2-5/i} \sim i^{-i-5/2} e^{-2}, \quad (\text{A10})$$

the previous result simplifies to

$$w_i = \frac{12^i}{\sqrt{\pi}} i^{-5/2}. \quad (\text{A11})$$

The corresponding selection bias $W(\mathbf{n})$ on arbitrary distribution $\mathbf{n} = (n_1, n_2, \dots)$ is

$$W(\mathbf{n}) = \prod_i w_i^{n_i} = \pi^{-N/2} 12^M \prod_i i^{-5n_i/2}, \quad (\text{A12})$$

where we have used $\sum n_i = N$ and $\sum i n_i = M$. The constant factors raised to N or M are selection-neutral (they make identical contribution to all distributions of the ensemble) and may be dropped from the selection bias so that the Stockmayer bias reduces to

$$w_i = i^{-5/2}. \quad (\text{A13})$$

The expressions for the cluster bias in Eqs. (A11) and (A13) are functionally equivalent, i.e., they produce identical ensembles.

-
- [1] J. W. Gibbs, *Elementary Principles in Statistical Mechanics* (Ox Bow Press, Woodbridge, CT, 1981), (reprint of the 1902 edition).
 - [2] R. D. Vigil, *Journal of Colloid and Interface Science* **336**, 642 (2009), URL <http://www.sciencedirect.com/science/article/pii/S002197970900455X>.
 - [3] N. Berestycki and J. Pitman, *Journal of Statistical Physics* **127**, 381 (2007), URL <http://dx.doi.org/10.1007/s10955-006-9261-1>.
 - [4] E. M. Hendriks, *Z. Phys. B Condensed Matter* **57**, 307 (1985).
 - [5] E. M. Hendriks, J. L. Spouge, M. Eibl, and M. Schreckenberg, *Zeitschrift für Physik B Condensed Matter* **58**, 219 (1985), URL <http://dx.doi.org/10.1007/BF01309254>.
 - [6] J. L. Spouge, *Macromolecules* **16**, 121 (1983), URL <http://pubs.acs.org/doi/abs/10.1021/ma00235a024>.
 - [7] P. J. Flory, *Journal of the American Chemical Society* **63**, 3091 (1941), <http://pubs.acs.org/doi/pdf/10.1021/ja01856a062>, URL <http://pubs.acs.org/doi/abs/10.1021/ja01856a062>.
 - [8] D. J. Aldous, *Bernoulli* **5**, 3 (1999).
 - [9] A. A. Lushnikov, *Phys. Rev. Lett.* **93**, 198302 1 (2004).
 - [10] E. D. McGrady and R. M. Ziff, *Phys. Rev. Lett.* **58**, 892 (1987), URL <http://journals.aps.org/prl/abstract/10.1103/PhysRevLett.58.892>.
 - [11] R. Pastor-Satorras and A. Vespignani, *Phys. Rev. Lett.* **86**, 3200 (2001), URL <http://link.aps.org/doi/10.1103/PhysRevLett.86.3200>.
 - [12] M. E. J. Newman, *Phys. Rev. E* **66**, 016128 (2002), URL <http://link.aps.org/doi/10.1103/PhysRevE.66.016128>.
 - [13] P. L. Krapivsky, S. Redner, and E. Ben-Naim, *A Kinetic View of Statistical Physics* (Cambridge University Press, Cambridge, 2010).
 - [14] P. Erdős and A. Rényi, *Math. Inst. Hungar. Acad. Sci* **5**, 17 (1960).
 - [15] M. E. J. Newman, *Phys. Rev. Lett.* **103**, 058701 (2009), URL <http://link.aps.org/doi/10.1103/PhysRevLett.103.058701>.
 - [16] J. Gao, S. V. Buldyrev, H. E. Stanley, and S. Havlin, *Nat Phys* **8**, 40 (2012), URL <http://dx.doi.org/10.1038/nphys2180>.
 - [17] I. Breskin, J. Soriano, E. Moses, and T. Tlusty, *Phys. Rev. Lett.* **97**, 188102 (2006), URL <http://link.aps.org/doi/10.1103/PhysRevLett.97.188102>.
 - [18] E. T. Jaynes, *Phys. Rev.* **106**, 620 (1957).
 - [19] T. Matsoukas and Y. Lin, *Phys. Rev. E* **74**, 031122 (pages 9) (2006), URL <http://link.aps.org/abstract/PRE/v74/e031122>.
 - [20] I. M. Gelfand and S. V. Fromin, *Calculus of Variations* (Dover, Mineola, NY, 2000), (reprint of the 1963 edition).
 - [21] J. Pitman, *Combinatorial Stochastic Processes*, vol. 1875 (Springer, 2006).
 - [22] R. M. Ziff, M. H. Ernst, and E. M. Hendriks, *Journal of Physics A: Mathematical and General* **16**, 2293 (1983), URL <http://stacks.iop.org/0305-4470/16/2293>.
 - [23] W. H. Stockmayer, *The Journal of Chemical Physics* **11**, 45 (1943), URL <http://link.aip.org/link/?JCP/11/45/1>.